Remarks as Delivered by Counselor John Podesta
The White House/MIT "Big Data" Privacy Workshop
March 3, 2014

Introduction

Good morning. I'm sorry to be talking to you remotely. In my world, big data squared off against big snow, and big snow won. Secretary Pritzker is travelling from NYC and I hope she'll have better luck and be with you at lunch. And the Administration is well represented by our Deputy CTO Nicole Wong and the head of NTIA Larry Strickling.

I want to start by thanking President Reif, for joining the White House in this important exploration of the technologies driving the big data revolution. And I want to thank Danny Weitzner and Elizabeth Bruce in particular, not only for putting together this outstanding event, but for their ongoing contributions to the research in this area.

This workshop is the first in a series of events which the White House will be co-hosting with academic institutions across the country. So, it is a fitting time for me to provide some background about the 90-day White House study of big data and privacy, the process that we're currently undertaking, and what we hope to accomplish in the next several weeks.

Background

As many of you will recall, on January 17, the President spoke to the American people about how to keep us safe from terrorism in a changing world, and at the same time continue to uphold America's commitment to liberty and privacy that our values and our Constitution require. In that speech, he asked me to lead a comprehensive review of big data and privacy, recognizing that national security is not the only space where changes in technology are altering the landscape of how data is collected and used, and challenging traditional conceptions of privacy. So, one purpose of this study is to get a more holistic view of the state of the technology and the benefits and challenges that it brings. This Administration remains committed to an open, interoperable, secure and reliable internet – the fundamentals that have enabled innovation to flourish, drive markets and improve lives. We also recognize that ensuring the continued strength of internet requires applying our timeless privacy values to these new technologies, as we have throughout our history with each new mode of communication from the mail to the telephone to the social network.

We are undergoing a revolution in the way that information about our purchases, our conversations, our social networks, our movements, and even our physical identities are collected, stored, analyzed and used. The immense volume, diversity, velocity, and potential value of data will have profound implications for privacy, the economy, and public policy. The White House working group will consider all those issues, and specifically how the present and future state of these technologies might motivate us to re-visit our policies across a range of sectors.

There is a lot of buzz these days about "Big Data" – a lot of marketing-speak and pitch materials for VC funding. For purposes of the White House study, when we talk about "big data" we're referring to data sets that are so large, so diverse, or so complex that the conventional tools that would ordinarily be used to manage data simply don't work. Instead, deriving value from these data sets require a series of more sophisticated techniques, such as Hadoop , NoSQL,

MapReduce and machine learning. These techniques enable the discovery of insights from big data sets that were not previously possible.

There is no question that there is more data than ever before, and no sign that the trajectory is slowing its upward pace. In 2012, there were an estimated 2.4 billion global internet users. The amount of global digital information created and shared – from documents to photos to videos to tweets – grew 9x in five years to nearly 2 zettabytes in 2011 (a zettabyte is one trillion gigabytes). On Facebook, there are some 350 million photos uploaded and shared every day. On YouTube, 100 hours of video is uploaded every minute. And we are only in the very nascent stage of the "Internet of Things," where our appliances will communicate with each other and sensors may be nearly ubiquitous.

The value that can be generated by the use of big data is not hypothetical. The availability of large data sets, and the computing power to derive value from them, is creating new business models, enabling innovations to improve efficiency and performance in a variety of public and private sector settings, and making possible valuable data-driven insights that are measurably improving outcomes in areas from education to healthcare. For example, The Cancer Genome Atlas, an NIH-funded program, is using large genomic data sets to map the genetic changes in more than 20 cancer types. Their researchers have discovered that breast and ovarian cancers have genomic similarities that may have implications for treating these diseases.

With the exponential advance of these capabilities, we must make sure that our modes of protecting privacy – whether technological, regulatory or social – also keep pace. Now, it's certainly true that data analytics is an old science, dating to the late 1800s. In this study, we want to explore whether there is something truly new in the vast collection of data and lightning-speed analytics that are made possible by new technologies, computational strategies and cratering storage costs. My hope is that this inquiry will anticipate future technological trends to help us frame the key questions arising from the collection, availability, and use of big data — both for our government, and the nation as a whole – and develop a workplan to address them.

Today's conference – appropriately set at MIT which has been the cradle for so many game-changing technologies – is part of this 90-day endeavor, and is designed to provide a firm grounding in the current state of technologies and their likely trajectories.

The Administration's Big Data initiatives

It is important to note that the Administration is not starting from scratch when it comes to big data or privacy.

Since the earliest days of this Administration, the Federal Government has taken unprecedented steps to make government data more available to citizens, companies and innovators. Through the Data.gov platform which launched in 2009, users have been able to access thousands of government datasets about a wide range of topics. The Open Data Initiative and Executive Order that the President signed last year commits federal agencies to unlocking even more valuable data from the vaults of government in health, energy, education, public safety, finance and global development.

The natural outgrowth of this commitment to making large data sets available for public innovation is a broad commitment to the technologies that can harness these assets. In 2012, the Administration announced a $200 million commitment by 6 agencies to invest in big data projects. And just last fall, we showcased 28 public-private partnerships harnessing big data to enhance national priorities, including economic growth and job creation, education and health, energy and sustainability, public safety and national security, and global development. Indeed, one of those projects was launched from here as part of MIT's Big Data Initiative. We are pleased to be collaborating with CSAIL's Big Data Privacy Working Group, and we look forward to hearing from some of the researchers engaged in that project later today.

The United States can also be proud of its long history as a leader in information privacy, starting with the pioneering of the Fair Information Practice Principles in the 1970s. Those principles -- known as the "FIPPs" -- are the underpinnings of the Privacy Act of 1974 which articulates the rights of citizens and the obligations of government to protect personal information. The same principles have also become the globally-recognized foundation for privacy protection, adopted by the OECD and providing the framework for privacy regimes around the world. And President Obama, from early in his first term, has been working to advance protections for individual privacy in this new age of information technology.

Indeed, the Administration announced a groundbreaking privacy document in 2012, with the release of its consumer privacy blueprint, including the Consumer Privacy Bill of Rights. The blueprint refined the FIPPs to be more focused on consumers in terms they could understand in their own lives. It also re-framed the FIPPs to better accommodate the incredibly innovative online environment in which we all now live. While the document does not specifically use the term "big data," the blueprint recognized that significant data was being collected about individuals online, and that some data would be sensitive. It also assumed that this data could deliver significant value, if properly used, to individual consumers.

What we will be exploring in this study is whether the Consumer Privacy Bill of Rights fully addresses the changes that today we refer to as the "big data revolution" – recognizing that we may only be at the beginning of that revolution. What the President wants to explore, in part, is whether our existing privacy framework can accommodate these changes, or if there are new avenues for policy that we need to consider.

Have we fully considered the myriad ways in which this data revolution might create social value, and have we fully contemplated the risks that it might pose to our conceptions of individual privacy, personal freedom and government responsibility of data?

As we move from predicated analysis of data – that is, using data to find something we already know that we're looking for -- to non-predicated, or pattern-based, searches – using data to find patterns that reveal new insights -- I think we need to be conscious of the implications for individuals.

How should we think about individuals' sense of their identity when data reveals things they didn't even know about themselves? In this study, we want to explore the capabilities of big data analytics, but also the social and policy implications of that capability.

Our work is still in its early stages, but already we're learning important things about the current state of technology and its potential. For example, we recently met with some leaders in higher education to discuss the use of academic performance data to improve learning outcomes. There is some terrific research happening in this area, and it worth talking about in a bit more detail.

The Pittsburgh Science of Learning Center – an NSF-funded center that joins the disciplines of cognitive learning and computer science -- hosts the "DataShop", the world's preeminent central repository for data on the interactions between students and educational software and a suite of tools to analyze that data. In collaboration with private sector partners, they have made large-scale data sets available to develop learning models aimed at improving math, science and language curriculums for K-12 students. In one study, the researchers tested a new algebra curriculum for middle school to high school students that utilized education technologies for instruction and to measure performance.

As some of you may know, mathematics proficiency rates of students in the United States – while on the rise – are still far below what they should be and lag behind students in the top-scoring countries. While there is still more work to be done, the early results of these large-scale studies show significant gains – 8 percentile points more than usual -- an amount that is nearly double how much students learn from a typical algebra course.

Importantly, what this type of research underscores is that the use of educational technologies is improving the scope, scale and granularity of data we have about how kids learn. With that data, and applying cognitive and data science to the problem, we are better able to understand how to help our children move up the performance curve. We are gaining insights that were not previously possible, or maybe were dismissed for lack of concrete data, because of the new capability to capture student performance in detail and at scale in diverse, real-world school contexts.

Of course, there are also privacy implications to be considered when gathering and using this data. While the educators working with the students obviously knew how individual kids were doing on their tests, the researchers who developed those data-driven education tools only had de-identified data and deliberately decided not to collect the demographic data of their students. Now, given what we already know about effective education policy, demographic data might have been useful both in developing effective curriculum and addressing the needs of the individual student. But the researchers decided that collection of such information raised privacy and ethical concerns, and that they could make progress without that data.

We can see similar real-world benefits and similar privacy questions raised in a range of areas: from tracking electricity usage in a home to significantly bring down energy costs, to collecting individual location data in order to reduce traffic congestion. I believe we'll be hearing about some of these innovative uses today, including uses in education, genomics, and transportation.

This is the power of big data analytics that could unleash real human potential, and so a goal of this study is to look at where the federal government can play a role in supporting this type of work while continuing to protect personal privacy and other values.

So that is the context of this inquiry.

The Study
Now, let me just take a few moments to explain a bit more about the review, its scope, and what you can expect over the next 90 days.

In his speech, the President asked me to lead a comprehensive review of the way that "big data" will affect the way we live and work; the relationship between government and citizens; and how public and private sectors can spur innovation and maximize the opportunities and free flow of this information while minimizing the risks to privacy. I will be joined in this effort by Secretary of Commerce Penny Pritzker (who will be your lunchtime keynote later today), Secretary of Energy Ernie Moniz, the President's Science Advisor John Holdren, the President's Economic Advisor Jeff Zients and other senior government officials.

This is going to be a collaborative effort with four channels of engagement.

First, the President's Council of Advisors on Science and Technology (PCAST) is conducting a parallel study to explore in-depth the technological dimensions of the intersection of big data and privacy. Their report will feed into this broader effort and ensure a substantive grounding in the technologies at issue.

Second, our working group is consulting with a wide range of stakeholders. We have already met with privacy and civil liberties advocates, business leaders, policymakers, international partners, academics and several government agencies on the significance of and future for these technologies. In the next several weeks, we look forward to hearing from a broad range of private sector companies, particularly those who collect and use data to develop products and deliver services, whether by targeted advertising; improved medical treatment; financial services, and more.

We also will engage international audiences, including international regulators and officials, to help answer the President's charge that we consider "whether we can forge international norms on how to manage this data; and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security."

Third, this workshop kicks off a series of events that we are co-hosting around the country to convene stakeholders to discuss these very issues and questions. The next event will be on March 17, co-hosted with the Data & Society Research Institute and NYU, and will focus on the social, cultural and ethical implications of big data. Then, on April 1, we will co-host an event with the School of Information and Berkeley Center for Law & Technology at UC Berkeley, which will focus on the legal and policy issues raised by big data.

Finally, and perhaps most importantly, we want to engage the public. This is not a discussion that should be confined to Washington or academia. This is an issue of such importance, an array of technologies already so pervasive, that it requires public participation in the conversation about how we realize the great benefits of big data while protecting individual privacy and other values. To this end, this week I will be posting a video to the White House website that describes this inquiry and asks the public "what technologies are most transformative in your life?" and "which technologies give you pause?" We have also just initiated a process to receive written comments addressing these questions in even more depth. You can find both channels for providing your

input on the White House website, and we welcome your comments and ideas. All of these discussions will help to inform our study.

This study is fundamentally a scoping exercise. We are trying to get a full view of the landscape – the technologies at play, the uses by the government, industry and academia. Whether we want to examine the Administration's consumer privacy blueprint—including the Consumer Privacy Bill of Rights—and how its principles can be applied in this new landscape. That may prompt us to look harder at some of our existing policies, at our research agenda, or at specific sectors where great gains could be made by the use of big data.

When we complete our work, we expect to deliver to the President a report that anticipates future technological trends and frames the key questions that the collection, availability, and use of "big data" raise – both for our government, and the nation as a whole. It will help identify technological changes to watch, whether those technological changes are addressed by the U.S.'s current policy framework and highlight where further government action, funding, research and consideration may be required.

While we don't expect to answer all these questions, or produce a comprehensive new policy in 90 days, we expect this work to serve as the foundation for a robust and forward-looking plan of action.

This is a fascinating and complex area, so let me close by throwing out a few questions that we have been thinking about:

What is genuinely "new" about big data and what, if any, policies should be revisited because of those changes?

What business models do you think are most dependent, today, on big data? How will that change in, say, 5 years? 15?

What types of uses of big data could measurably improve social or economic outcomes or productivity with further government action, funding or research?

Can we "build in" additional privacy protection into the architecture of big data analytics and should the government and the private sector be investing more in research toward that end.

For individuals, what do you think will be the most significant effects of these emerging pattern-based data mining techniques?

Thank you for your time this morning and your engagement in this national conversation. I am sorry I am not with you in person, but I'll be watching the feed. If we can manage the technology, I think a have a few more minutes to take some questions.